Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 12:53:16
REFERENCE NO: 239

Author Names & Affiliations

- Matthew Browning - College of Applied Health Sciences, University of Illinois at Urbana-Champaign
- Roberto Aldunate - College of Applied Health Sciences, University of Illinois at Urbana-Champaign
- Ian Brooks - ISchool, University of Illinois at Urbana-Champaign
- Melissa Cragin - Midwest Big Data Hub

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Health informatics/sciences

Title of Submission

Advanced cyberinfrastructure for health care data

Abstract (maximum ~200 words).

Information about medical and public health problems at micro and macro levels are not readily and widely available for scientists, engineers, and other scholars interested in understanding and improving medical and public health systems nationwide. This document proposes to develop an advanced Health Claim Cyberinfrastructure, given the huge advantages that this type of data has over medical records to conduct data analyses at large scales. This approach would benefit not only institutions involved with these datasets, but also a large community of individual PIs. They would be able to conduct pilot analysis in order to explore/discover patterns and emerging trends in health care provisioning nationwide, simulate the impact of policies, understand boundary conditions for such policies, and ultimately help the fine-tuning of medical and public health systems. In addition, this advanced Health Claim Cyberinfrastructure would provide a unique opportunity to generate efficiencies at scale for topics such as HIPAA and accessibility.

Question 1 Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

Researchers cannot readily access the rich sets of data that could help answer 21st century medical and public health problems. These rich datasets are housed in health insurance claims databases both at private companies and federal programs (Medicare, Medicaid). Much of these data are completely blocked to outside researchers and only used by in-house market research teams. Other datasets are de-identified and made accessible to a lucky few academic institutions. Providing greater access to these data would accelerate health-related science and engineering research and have great impacts on health care policy and practice.

Claims data have huge advantages over electronic medical records from individual providers - the norm in epidemiological public health studies. People tend to subscribe to insurance plans for years or decades but switch providers based on their specific health care needs and procedures. As such, medical records are often constrained to types and timings of services. In contrast, claims data provide subscriber-level coverage regardless of where and when they saw a doctor, bought a medication, or received in-home care. Claims data also have records of health care expenditures, including inpatient, outpatient, laboratory, and pharmacy costs. Given the exorbitant cost of health care today, claims data are increasingly important for big data research and healthcare policy implications.

Currently, claims data are only available through difficult-to-navigate pathways, making many such datasets impossible for most academic institutions to use. These navigational pathways are very costly, both in regards to time and money. University negotiators become tied up in what are often one-off agreements. Lack of access not only disadvantages faculty at non-R1 universities (which are often particularly limited in their ability to navigate these pathways) but also limits scholarship because of the closed sets of ideas that get brought to bear.

One national non-profit organization, the Health Care Cost Initiative (http://www.healthcostinstitute.org) has consolidated Aetna, Humana, Kaiser Permanente, and UnitedHealthcare claims, but accessing HCCI data requires a university-level financial commitment which blocks individual PIs from being able to apply for access. Furthermore, HCCI does not allow pilot studies, and this situation further limits PI's ability to write competitive funding applications to access these data. Despite these obstacles to HCCI data, these datasets are extremely high in demand so much so that new data use agreements are only allowed every three years and fill up in a matter of days. A few other organizations and companies provide health insurance data. OptumHealth provides access to UnitedHealthcare claims, but the financial investments required to work with Optum again require university-level commitment and are impossible for individual PIs to access. Nation-wide Blue Cross Blue Shield claims recently became available to a few elite universities in an August 2016 pilot program. Based on our experience, only Kaiser Permanente allows partnerships with individual PIs. Projects that are of mutual interest to both KP and university parties may allow PIs to even access geo-coded individual records. However, PIs can only consider accessing KP data if they know these resources exist and connect with the right people to access these data.

Meanwhile, federal claims (such as Medicare) require complicated data use and sharing agreements between several parties including the National Institutes of Health, National Bureau for Economic Research, and Centers for Medicare and Medicaid Services. These processes must be repeated each time such data are requested. This again makes it difficult for individual PIs to access data. Moreover, these processes can be lengthy and difficult for PIs who have not already undertaken this process.

The cumulative result of this situation in which insurance claims are overly difficult to access is that researchers, particularly at the junior level, are stymied and unable to study the Grand Challenges that such data could address. We propose that NSF invest in developing cyberinfrastructure policies and data use pathways that facilitate access and computing resources to health insurance claims datasets. Such investments would help answer many of the questions that researchers today cannot adequately answer with the public health datasets readily available.

Question 2 Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified research challenge(s).

This document proposes the development of a Health Claims Cyberinfrastructure (HCC), where the different stakeholders would use claims datasets and the health cyberinfrastructure coordination and information sharing mechanisms to improve health services provided to the population. Simultaneously, HCC would enable the scientific community to tackle challenging problems that can not be approached at smaller scales, such as trends, global market analysis, non-linear dynamics, national policy development, among others.

The Health Claims Cyberinfrastructure proposed in this document would also work as a hub to streamline HIPAA and corporate proprietary-sensitive data access, which will allow universities, research centers, industry, insurance companies, and government agencies sharing data to converge into efficient and effective mechanisms for data gathering, data storage, data management, data integration, and data analysis.

A particular feature for the cyberinfrastructure needed to approach the research challenges described in the previous sections is the design of HCC to satisfy multi-scale memberships, such as institutional and PI-level access to HCC. This would greatly help facilitate access to

health insurance claims datasets for PIs to conduct pilot studies. This process would then have the potential to generate a quantum leap in health related research and innovation, as well as creating and strengthening relationships and collaborations between researchers and multi-disciplinary research and development initiatives.

In terms of implementation, HCC could leverage resources available and created by XSEDE (Extreme Science and Engineering Discovery Environment), led by the National Center for Supercomputer Applications (NCSA) from the University of Illinois at Urbana-Champaign. XSEDE could contribute to HCC as it is a single virtual system that scientists can use to interactively share computing resources, data, and expertise. This approach would also save time and resources in developing infrastructure services, such as efficient data storage and data sharing mechanisms. Additional resources would be necessary to be created by HCC, such as simulation tools for large health datasets, multi-campus/multi-institution models for sensitive data handling and sharing, and the integration of government, health industry/corporations, health insurance, and research institutions.

Encouraging data providers to contribute to HCC would require an understanding of the provider's needs and limitations of internal research operations. One of HCC's goals would be to facilitate provision of valuable services to data providers - not including market or competitive research. To determine ways in which HCC could benefit data providers, focus groups with stakeholders from pre-existing partnerships with data providers should be held. The Midwest is home to partnerships appropriate for this research. Several universities have partnered with regional health providers which have their own built-in insurance services and therefore would be excellent groups from which to elicit information (i.e., the University of South Dakota and Sanford Health partnership; the Wayne State University and Henry Ford Health System partnership; and the University of Illinois at Urbana-Champaign and Carle and Health Alliance Insurance partnerships).

In order to reduce the risk of development and to improve the soundness of the HCC with the stakeholders' community, it is envisioned that a spiral methodology would be used for this purpose. A local/regional model of implementation would be necessary for it to provide insight on strengths in development of the HCC and also to help discover the limitations of HCC. Next, a larger scope model would be phased to include the larger Midwest region, then to finally expand it to national levels.

Specific components of the HCC should include computing systems, data storage systems, advanced programs and repositories, metadata schemes, visualization environments, and people. While the justification for computing systems, data storage, and repositories may be straightforward, metadata and visualization are not. Metadata schemes are necessary to enable integration of datasets coming from different health insurance companies and organizations; hence the HCC would require developing a national standard for health care claim datasets. These metadata schemes should be flexible enough to not only satisfy effective and high-performance data sharing but also to allow HIPAA standards as needed. In terms of visualization, HCC should include as much data analytics as possible and aim for software re-use by allowing, facilitating, and promoting researchers to collectively and permanently build data analysis and visualization libraries available to the entire community, based on their profiles/roles.

Finally, a very important aspect for the College of Applied Health Sciences at the University of Illinois at Urbana-Champaign is the need that the proposed HCC satisfies: accessibility standards that allow any person with disabilities to access and use the data available at HCC in the same way as would a person without such disabilities. This component of the proposed HCC would require close collaboration with government agencies to fit Section 508 of the Rehabilitation Act of 1973. This capability is not only important for the HCC itself but would also be a valuable experience for deriving knowledge, guidelines, and lessons to contribute to the development of accessibility for existing and future cyberinfrastructure efforts.

Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode)."